PB-0008-1CIP

DIAGNOSTICS AND THERAPEUTICS FOR PANCREATIC DISORDERS

This application is a continuation-in-part of USSN 09/226,994, filed 7 January 1999.

FIELD OF THE INVENTION

The invention relates todiscovery of thirteen isolated polynucleotides and their encoded proteins that are highly co-expressed with genes known to be involved in insulin synthesis and useful for diagnosis, prognosis, and treatment of pancreatic disorders.

BACKGROUND OF THE INVENTION

Insulin is a hormone produced in the beta islet cells of the pancreas. Patients with diabetes have serum glucose levels that are chronically elevated above normal because they either produce insufficient insulin (type I diabetes) or are resistant to insulin (type II diabetes). Complications of diabetes include angina, hypertension, myocardial infarctions, peripheral vascular disease, diabetic retinopathy, diabetic nephropathy, diabetic necrosis, ulceration, and diabetic neuropathy (Davidson (1998) Diabetes Mellitus, WB Saunders, Philadelphia PA).

While some genes that participate in or regulate insulin synthesis and release are known, many genes that function in these critical pathways remain to be identified. Identification of currently unknown genes will provide surrogate diagnostic markers and new therapeutic targets.

Thus the present invention satisfies a need in the art by providing new compositions that are useful for diagnosis, prognosis, treatment, and evaluation of therapies for pancreatic disorders, especially diabetes. A method for analyzing gene expression patterns has been used to identity thirteen polynucleotides that have highly significant co-expression with genes known to be involved with insulin-synthesis.

SUMMARY OF THE INVENTION

The invention provides a composition comprising a plurality of polynucleotides having the nucleic acid sequences of SEQ ID NOs:1-13 or the complements thereof that are highly significantly co-expressed with genes such as insulin, glucagon, lipase, colipase, human islet amyloid polypeptide (HiAPP) and Reg-1 alpha, Reg-1 beta, and Reg-related regenerating genes (Reg), known to involved in insulin synthesis. The invention also provides an isolated polynucleotide comprising a nucleic acid sequence selected from SEQ ID NOs:1-13 or the complement thereof. In different aspects, the polynucleotide is used as a surrogate marker, as a probe, in an expression vector, and in the diagnosis, prognosis, evaluation of therapies and treatment of pancreatic disorders. The invention further provides a composition comprising a polynucleotide and a labeling moiety.

The invention provides a method for using a composition or a polynucleotide of the invention to screen a plurality of molecules and compounds to identify ligands which specifically bind to the composition or the polynucleotide. The molecules are selected from DNA molecules, RNA molecules, peptide nucleic acids,

peptides, mimetics, ribozymes, transcription factors, enhancers, and repressors. The invention also provides a method of using a composition or a polynucleotide to purify a ligand.

The invention provides a method for using a composition or an isolated polynucleotide to detect gene expression in a sample by hybridizing the composition or polynucleotide to nucleic acids of the sample under conditions for formation of one or more hybridization complexes and detecting hybridization complex formation, wherein complex formation indicates gene expression in the sample. In one aspect, the composition or polynucleotide is attached to a substrate. In another aspect, the nucleic acids of the sample are amplified prior to hybridization. In yet another aspect, complex formation is compared with at least one standard and indicates the presence of a pancreatic disorder.

The invention provides a purified protein or a portion thereof selected from SEQ ID NOs:14 and 15, which is encoded by a polynucleotide that is highly significantly co-expressed with genes known to involved in insulin synthesis and whose expression is associated with pancreatic disorders. The invention also provides a method for using a protein to screen a plurality of molecules to identify at least one ligand which specifically binds the protein. The molecules are selected from aptamers, DNA molecules, RNA molecules, peptide nucleic acids, peptides, mimetics, ribozymes, proteins, antibodies, agonists, antagonists, immunoglobulins, inhibitors, pharmaceutical agents or drug compounds. The invention further provides a method of using a protein to purify a ligand.

The invention provides a method of using a protein to make an antibody that specifically binds to the protein of the invention, and methods for using the antibody to diagnose or treat a pancreatic disorder. The invention also provides a composition comprising a polynucleotide, a protein, or an antibody that specifically binds a protein and a pharmaceutical carrier.

BRIEF DESCRIPTION OF THE SEQUENCE LISTING

The Sequence Listing provides exemplary polynucleotides comprising the nucleic acid sequences of SEQ ID NOs:1-13 some of which encode the proteins comprising the amino acid sequences of SEQ ID NOs:14 and 15. Each sequence is identified by a sequence identification number (SEQ ID NO) and by the Incyte clone number with which the sequence was first identified.

DESCRIPTION OF THE INVENTION

It must be noted that as used herein and in the appended claims, the singular forms "a", "an", and "the" include the plural reference unless the context clearly dictates otherwise. Thus, for example, a reference to "a host cell" includes a plurality of such host cells, and a reference to "an antibody" is a reference to one or more antibodies and equivalents thereof known to those skilled in the art, and so forth.

DEFINITIONS

"Markers for pancreatic disorders" refers to polynucleotides, proteins, and antibodies which are useful in the diagnosis, prognosis, evaluation of therapies and treatment of pancreatic disorders. Typically, this means that the marker gene is differentially expressed in samples from subjects predisposed to, manifesting, or diagnosed with a pancreatic disorder.

"Differential expression" refers to an increased or up-regulated or a decreased or down-regulated expression as detected by presence, absence or at least about a two-fold change in the amount of transcribed messenger RNA or protein in a sample.

"Pancreatic disorders" specifically include, but are not limited to, the following conditions, diseases, and disorders: type I and type II diabetes; complications of diabetes including angina, hypertension, myocardial infarctions, peripheral vascular disease, diabetic retinopathy, diabetic nephropathy, diabetic necrosis, ulceration, and diabetic neuropathy; islet cell hyperplasia; pancreatitis; and pancreatic tumor.

"Isolated or purified" refers to a polynucleotide or protein that is removed from its natural environment and that is separated from other components with which it is naturally present.

"Genes known to be highly expressed in insulin synthesis pathways" which were used in the co-expression analysis included insulin, glucagon, lipase, colipase, human islet amyloid polypeptide (HiAPP) and Reg-1 alpha, Reg-1 beta, and Reg-related regenerating genes (Reg).

"Polynucleotide" refers to an isolated cDNA. It can be of genomic or synthetic origin, double-stranded or single-stranded, and combined with vitamins, minerals, carbohydrates, lipids, proteins, or other nucleic acids to perform a particular activity or form a useful composition.

"Protein" refers to a purified polypeptide whether naturally occurring or synthetic.

"Sample" is used in its broadest sense. A sample containing nucleic acids can comprise a bodily fluid; an extract from a cell; a chromosome, organelle, or membrane isolated from a cell; genomic DNA, RNA, or cDNA in solution or bound to a substrate; a cell; a tissue; a tissue print; and the like.

"Substrate" refers to any rigid or semi-rigid support to which polynucleotides or proteins are bound and includes membranes, filters, chips, slides, wafers, fibers, magnetic or nonmagnetic beads, gels, capillaries or other tubing, plates, polymers, and microparticles with a variety of surface forms including wells, trenches, pins, channels and pores.

A "transcript image" is a profile of gene transcription activity in a particular tissue at a particular time.

A "variant" refers to a polynucleotide or protein whose sequence diverges from about 5% to about 30% from the nucleic acid or amino acid sequences of the Sequence Listing.

THE INVENTION

The present invention employed "guilt by association or GBA", a method for using marker genes known

to be associated with a particular condition, disease or disorder to identify surrogate markers, polynucleotides and their encoded proteins, that are similarly associated or co-expressed in the same condition, disease, or disorder (Walker and Volkmuth (1999) Prediction of gene function by genome-scale expression analysis: prostate-associated genes. Genome Res 9:1198-1203, incorporated herein by reference). In particular, the method identifies cDNAs cloned from mRNA transcripts which are active in tissues known to have been removed from subjects with pancreatic disorders. The polynucleotides, their encoded proteins and antibodies which specifically bind to the encoded proteins are useful for diagnosis, prognosis, evaluation of therapies, and treatment of pancreatic disorders.

Guilt by association provides for the identification of polynucleotides that are expressed in a plurality of libraries. The polynucleotides represent genes of unknown function which are expressed in a specific signaling pathway, disease process, subcellular compartment, cell type, tissue, or species. The expression patterns of the genes known to be highly expressed during insulin synthesis; insulin, glucagon, lipase, colipase, HiAPP, and Reg; are compared with those of polynucleotides with unknown function to determine whether a specified co-expression probability threshold is met. Through this comparison, a subset of the polynucleotides having a high co-expression probability with the known marker genes can be identified.

The polynucleotides originate from human cDNA libraries. These polynucleotides can also be selected from a variety of sequence types including, but not limited to, expressed sequence tags (ESTs), assembled polynucleotides, full length coding regions, and 3' untranslated regions. To be considered in GBA or co-expression analysis, the polynucleotides had to have been expressed in at least five cDNA libraries. In this application, GBA was applied to a total of 41,419 assembled polynucleotide bins that met the criteria of having been expressed in at least five libraries.

The cDNA libraries used in the co-expression analysis were obtained from adrenal gland, biliary tract, bladder, blood cells, blood vessels, bone marrow, brain, bronchus, cartilage, chromaffin system, colon, connective tissue, cultured cells, embryonic stem cells, endocrine glands, epithelium, esophagus, fetus, ganglia, heart, hypothalamus, hemic/immune system, intestine, islets of Langerhans, kidney, larynx, liver, lung, lymph, muscles, neurons, ovary, pancreas, penis, phagocytes, pituitary, placenta, pleura, prostate, salivary glands, seminal vesicles, skeleton, spleen, stomach, testis, thymus, tongue, ureter, uterus, and the like. The number of cDNA libraries analyzed can range from as few as three to greater than 10,000 and preferably, the number of the cDNA libraries is greater than 500.

In a preferred embodiment, the polynucleotides are assembled from related sequences, such as sequence fragments derived from a single transcript. Assembly of the polynucleotide can be performed using

sequences of various types including, but not limited to, ESTs, extension of the EST, shotgun sequences from a cloned insert, or full length cDNAs. In a most preferred embodiment, the polynucleotides are derived from human sequences that have been assembled using the algorithm disclosed in USSN 9,276,534, filed March 25, 1999, and used in USSN 09/226,994, filed 7 January 1999, both incorporated herein by reference.

Experimentally, differential expression of the polynucleotides can be evaluated by methods including, but not limited to, differential display by spatial immobilization or by gel electrophoresis, genome mismatch scanning, representational difference analysis, and transcript imaging. The results of transcript imaging for SEQ ID NO:2 are shown in Example IX . Differential expression of SEQ ID NO:2 is highly specifically correlated with type I diabetes. The transcript image provided direct confirmation of the strength of co-expression analysis--the use of known genes to identify unknown polynucleotides and their encoded proteins which are highly significantly associated with insulin synthesis and pancreatic disorders. Additionally, differential expression can be assessed by microarray technology. These methods can be used alone or in combination.

Genes known to be highly expressed in pancreatic disorders can be selected based on research in which the genes are found to be key elements of insulin synthesis pathways or on the known use of the genes as diagnostic or prognostic markers or therapeutic targets for pancreatic disorders. Preferably, the known genes are insulin, glucagon, lipase, colipase, HiAPP, and Reg.

The procedure for identifying novel polynucleotides that exhibit a statistically significant co-expression pattern with known genes is as follows. First, the presence or absence of a polynucleotide in a cDNA library is defined: a polynucleotide is present in a cDNA library when at least one cDNA fragment corresponding to the polynucleotide is detected in a cDNA from that library, and a polynucleotide is absent from a library when no corresponding cDNA fragment is detected.

Second, the significance of co-expression is evaluated using a probability method to measure a due-to-chance probability of the co-expression. The probability method can be the Fisher exact test, the chi-squared test, or the kappa test. These tests and examples of their applications are well known in the art and can be found in standard statistics texts (Agresti (1990) Categorical Data Analysis, John Wiley & Sons, New York NY; Rice (1988) Mathematical Statistics and Data Analysis, Duxbury Press, Pacific Grove CA). A Bonferroni correction (Rice, supra, p. 384) can also be applied in combination with one of the probability methods for correcting statistical results of one polynucleotide versus multiple other polynucleotides. In a preferred embodiment, the due-to-chance probability is measured by a Fisher exact test, and the threshold of the due-to-chance probability is set preferably to less than 0.001, more preferably to less than 0.00001.

For example, to determine whether two genes, A and B, have similar co-expression patterns,

occurrence data vectors can be generated as illustrated in the table below. The presence of a gene occurring at least once in a library is indicated by a one, and its absence from the library, by a zero.

	Library 1	Library 2	Library 3	...	Library N
Gene A	1	1	0	...	0
Gene B	1	0	1	...	0

For a given pair of genes, the occurrence data in the table above can be summarized in a 2 x 2 contingency table. The second table (below) presents co-occurrence data for gene A and gene B in a total of 30 libraries. Both gene A and gene B occur 10 times in the libraries.

	Gene A Present	Gene A Absent	Total
Gene B Present	8	2	10
Gene B Absent	2	18	20
Total	**10**	**20**	**30**

The second table summarizes and presents: 1) the number of times gene A and B are both present in a library; 2) the number of times gene A and B are both absent in a library; 3) the number of times gene A is present, and gene B is absent; and 4) the number of times gene B is present, and gene A is absent. The upper left entry is the number of times the two genes co-occur in a library, and the middle right entry is the number of times neither gene occurs in a library. The off diagonal entries are the number of times one gene occurs, and the other does not. Both A and B are present eight times and absent 18 times. Gene A is present, and gene B is absent, two times; and gene B is present, and gene A is absent, two times. The probability ("p-value") that the above association occurs due to chance as calculated using a Fisher exact test is 0.0003.

This method of estimating the probability for co-expression makes several assumptions. The method assumes that the libraries are independent and are identically sampled. However, in practical situations, the selected cDNA libraries are not entirely independent, because more than one library can be obtained from a single subject or tissue. Nor are they entirely identically sampled, because different numbers of cDNAs can have been sequenced from each library. The number of cDNAs sequenced typically ranges from 5,000 to 10,000 cDNAs per library. After the Fisher exact co-expression probability is calculated for each polynucleotide versus all other assembled polynucleotides that occur, a Bonferroni correction for multiple statistical tests is applied.

Using the method of the present invention, we have identified polynucleotides, SEQ ID NOs:1-13 and their encoded proteins, SEQ ID NOs:14 and 15, that exhibit highly significant co-expression probability with known marker genes for pancreatic disorders. The results presented in Example VI show the direct (known gene to unknown polynucleotide) or indirect (known gene to unknown polynucleotide to a second unknown polynucleotide) associations among the novel polynucleotides and the known marker genes for pancreatic disorders. Therefore, by these associations, the novel polynucleotides are useful as surrogate markers for the co-expressed known marker genes in diagnosis, prognosis, evaluation of therapies and treatment of pancreatic disorders. Further, the proteins or peptides expressed from the novel polynucleotides are either potential therapeutics or targets for the identification and/or development of therapeutics.

In one embodiment, the present invention encompasses a composition comprising a plurality of polynucleotides having the nucleic acid sequences of SEQ ID NOs:1-13 or the complements thereof. These thirteen polynucleotides are shown by the method of the present invention to have significant co-expression with known genes associated with pancreatic disorders. The invention also provides a polynucleotide, its complement, a probe comprising the polynucleotide or the complement thereof selected from SEQ ID NOs:1-13 and variants thereof.

The polynucleotide can be used to search against the GenBank primate (pri), rodent (rod), mammalian (mam), vertebrate (vrtp), and eukaryote (eukp) databases; the encoded protein, against GenPept, SwissProt, BLOCKS (Bairoch et al. (1997) Nucleic Acids Res 25:217-221), PFAM, and other databases that contain previously identified and annotated protein sequences, motifs, and gene functions. Methods that search for primary sequence patterns with secondary structure gap penalties (Smith et al. (1992) Protein Engineering 5:35-51) as well as algorithms such as Basic Local Alignment Search Tool (BLAST; Altschul (1993) J Mol Evol 36:290-300; Altschul et al. (1990) J Mol Biol 215:403-410), BLOCKS (Henikoff and Henikoff (1991) Nucleic Acids Res 19:6565-6572), Hidden Markov Models (HMM; Eddy (1996) Cur Opin Str Biol 6:361-365; Sonnhammer et al. (1997) Proteins 28:405-420), and the like, can be used to manipulate and analyze nucleotide and amino acid sequences. These databases, algorithms and other methods are well known in the art and are described in Ausubel et al. (1997; Short Protocols in Molecular Biology, John Wiley & Sons, New York NY, unit 7.7) and in Meyers (1995; Molecular Biology and Biotechnology, Wiley VCH, New York NY, p 856-853).

Also encompassed by the invention are polynucleotides that are capable of hybridizing to SEQ ID NOs:1-13 and the complements thereof under highly stringent conditions. Stringency can be defined by salt concentration, temperature, and other chemicals and conditions well known in the art. Conditions can be selected, for example, by varying the concentrations of salt in the prehybridization, hybridization, and wash

solutions or by varying the hybridization and wash temperatures. With some substrates, the temperature can be decreased by adding a solvent such as formamide to the prehybridization and hybridization solutions.

Hybridization can be performed at low stringency, with buffers such as 5xSSC (saline sodium citrate) with 1% sodium dodecyl sulfate (SDS) at 60C, which permits complex formation between two nucleic acid sequences that contain some mismatches. Subsequent washes are performed at higher stringency with buffers such as 0.2xSSC with 0.1% SDS at either 45C (medium stringency) or 68C (high stringency), to maintain hybridization of only those complexes that contain completely complementary sequences. Background signals can be reduced by the use of detergents such as SDS, sarcosyl, or TRITON X-100 (Sigma-Aldrich, St. Louis MO), and/or a blocking agent, such as salmon sperm DNA. Hybridization methods are described in detail in Ausubel (supra, units 2.8-2.11, 3.18-3.19 and 4-6-4.9) and Sambrook et al. (1989; Molecular Cloning, A Laboratory Manual, Cold Spring Harbor Press, Plainview NY).

A polynucleotide can be extended utilizing primers and employing various PCR-based methods known in the art to detect upstream sequences such as promoters and other regulatory elements. (See, e.g., Dieffenbach and Dveksler (1995) PCR Primer, a Laboratory Manual, Cold Spring Harbor Press, Plainview NY.) Commercially available kits such as XL-PCR (Applied Biosystems, Foster City CA), cDNA libraries (Life Technologies, Rockville MD) or genomic libraries (Clontech, Palo Alto CA) and nested primers can be used to extend the sequence. For all PCR-based methods, primers can be designed using commercially available software (LASERGENE software, DNASTAR, Madison WI) or another program, to be about 15 to 30 nucleotides in length, to have a GC content of about 50%, and to form a hybridization complex at temperatures of about 68C to 72C.

In another aspect of the invention, the polynucleotide can be cloned into a recombinant vector that directs the expression of the protein, or structural or functional portions thereof, in host cells. Due to the inherent degeneracy of the genetic code, other DNA sequences which encode functionally equivalent amino acid sequence can be produced and used to express the protein encoded by the polynucleotide. The nucleotide sequences of the present invention can be engineered using methods generally known in the art in order to alter the nucleotide sequences for a variety of purposes including, but not limited to, modification of the cloning, processing, and/or expression of the gene product. DNA shuffling by random fragmentation, as described in USPN 5,830,721, and PCR reassembly of gene fragments and synthetic oligonucleotides can be used to engineer the nucleotide sequences. For example. oligonucleotide-mediated site-directed mutagenesis can be used to introduce mutations that create new restriction sites, alter glycosylation patterns, change codon preference, produce splice variants, and so forth.

In order to express a biologically active protein, the polynucleotide or derivatives thereof, can be inserted into an expression vector with elements for transcriptional and translational control of the inserted coding sequence in a particular host. These elements include regulatory sequences, such as enhancers, constitutive and inducible promoters, and 5' and 3' untranslated regions. Methods which are well known to those skilled in the art can be used to construct such expression vectors. These methods include in vitro recombinant DNA techniques, synthetic techniques, and in vivo genetic recombination (Ausubel, supra, unit 16).

A variety of expression vector/host cell systems can be utilized to express the polynucleotide. These include, but are not limited to, microorganisms such as bacteria transformed with recombinant bacteriophage, plasmid, or cosmid expression vectors; yeast transformed with yeast expression vectors; insect cell systems infected with baculovirus vectors; plant cell systems transformed with viral or bacterial expression vectors; or animal cell systems. For long term production of recombinant proteins in mammalian systems, stable expression in cell lines is preferred. For example, the polynucleotide can be transformed into cell lines using expression vectors which can contain viral origins of replication and/or endogenous expression elements and a selectable or visible marker gene on the same or on a separate vector. The invention is not to be limited by the vector or host cell employed.

In general, host cells that contain the polynucleotide and that express the protein can be identified by a variety of procedures known to those of skill in the art. These procedures include, but are not limited to, DNA-DNA or DNA-RNA hybridizations, PCR amplification, and protein bioassay or immunoassay techniques which include membrane, solution, or chip-based technologies for the detection and/or quantification of nucleic acid or amino acid sequences. Immunological methods for detecting and measuring the expression of the protein using either specific polyclonal or monoclonal antibodies are known in the art. Examples of such techniques include enzyme-linked immunosorbent assays (ELISAs), radioimmunoassays (RIAs), and fluorescence activated cell sorting (FACS).

Host cells transformed with the polynucleotide can be cultured under conditions for the expression and recovery of the protein from cell culture. The protein produced by a transgenic cell can be secreted or retained intracellularly depending on the sequence and/or the vector used. As will be understood by those of skill in the art, expression vectors containing the polynucleotide can be designed to contain signal sequences which direct secretion of the protein through a prokaryotic cell wall or eukaryotic cell membrane.

In addition, a host cell strain can be chosen for its ability to modulate expression of the inserted sequences or to process the expressed protein in the desired fashion. Such modifications of the protein include, but are not limited to, acetylation, carboxylation, glycosylation, phosphorylation, lipidation, and acylation. Post-

translational processing which cleaves a "prepro" form of the protein can also be used to specify protein targeting, folding, and/or activity. Different host cells which have specific cellular machinery and characteristic mechanisms for post-translational activities (e.g., CHO, HeLa, MDCK, HEK293, and WI38) are available from the ATCC (Manassas VA) and can be chosen to ensure the correct modification and processing of the expressed protein.

In another embodiment of the invention, natural, modified, or recombinant polynucleotides are ligated to a heterologous sequence resulting in translation of a fusion protein containing heterologous protein moieties in any of the aforementioned host systems. Such heterologous protein moieties facilitate purification of fusion proteins using commercially available affinity matrices. Such moieties include, but are not limited to, glutathione S-transferase, maltose binding protein, thioredoxin, calmodulin binding peptide, 6-His, FLAG, c-myc, hemaglutinin, and monoclonal antibody epitopes.

In another embodiment, the polynucleotides, wholly or in part, are synthesized using chemical or enzymatic methods well known in the art (Caruthers et al. (1980) Nucl Acids Symp Ser (7) 215-233; Ausubel, supra, units 10.4 and 10.16). Peptide synthesis can be performed using various solid-phase techniques (Roberge et al. (1995) Science 269:202-204), and machines such as the ABI 431A peptide synthesizer (Applied Biosystems) can be used to automate synthesis. If desired, the amino acid sequence can be altered during synthesis to produce a more stable variant for therapeutic use.

SCREENING, DIAGNOSTICS AND THERAPEUTICS

The polynucleotides can be used as surrogate markers in diagnosis, prognosis, evaluation of therapies and treatment of pancreatic disorders including, but not limited to, type I and type II diabetes; complications of diabetes including angina, hypertension, myocardial infarctions, peripheral vascular disease, diabetic retinopathy, diabetic nephropathy, diabetic necrosis, ulceration, and diabetic neuropathy; islet cell hyperplasia; pancreatitis; and pancreatic tumor.

The polynucleotide can be used to screen a plurality or library of molecules and compounds for specific binding affinity. The assay can be used to screen DNA molecules, RNA molecules, peptide nucleic acids, peptides, mimetics, ribozymes, or proteins including transcription factors, enhancers, repressors, and the like which regulate the activity of the polynucleotide in the biological system. The assay involves providing a plurality of molecules and compounds, combining a polynucleotide or a composition of the invention with the plurality of molecules and compounds under conditions to allow specific binding, and detecting specific binding to identify at least one molecule or compound which specifically binds at least one polynucleotides of the invention.

Similarly the proteins, or portions thereof, can be used to screen a plurality or library of molecules or compounds in any of a variety of screening assays to identify a ligand. The protein employed in such screening

can be free in solution, affixed to an abiotic substrate or expressed on the external, or a particular internal surface, of a bacterial, or other, cell. Specific binding between the protein and the ligand can be measured. The assay can be used to screen aptamers, DNA molecules, RNA molecules, peptide nucleic acids, peptides, mimetics, ribozymes, proteins, antibodies, agonists, antagonists, immunoglobulins, inhibitors, pharmaceutical agents or drug compounds and the like, which specifically bind the protein. One method for high throughput screening using very small assay volumes and very small amounts of test compound is described in Burbaum et al. USPN 5,876,946, incorporated herein by reference, which screens large numbers of molecules for enzyme inhibition or receptor binding.

In one preferred embodiment, the polynucleotides are used for diagnostic purposes to determine the differential expression of a gene in a sample. The polynucleotide consists of complementary RNA and DNA molecules, branched nucleic acids, and/or PNAs. In one alternative, the polynucleotides are used to detect and quantify gene expression in biopsied samples in which differential expression of the polynucleotide indicates the presence of a disorder. In another alternative, the polynucleotide can be used to detect genetic polymorphisms associated with a disease or disorder. In a preferred embodiment, these polymorphisms are detected in an mRNA transcribed from an endogenous gene.

In another preferred embodiment, the polynucleotide is used as a probe. Specificity of the probe is determined by whether it is made from a unique region, a regulatory region, or from a region encoding a conserved motif. Both probe specificity and the stringency of the diagnostic hybridization or amplification will determine whether the probe identifies only naturally occurring, exactly complementary sequences, allelic variants, or related sequences. Probes designed to detect related sequences should preferably have at least 50% sequence identity to at least a fragment of a polynucleotide of the invention.

Methods for producing hybridization probes include the cloning of nucleic acid sequences into vectors for the production of RNA probes. Such vectors are known in the art, are commercially available, and can be used to synthesize RNA probes in vitro by adding RNA polymerases and labeled nucleotides. Probes can incorporate nucleotides labeled by a variety of reporter groups including, but not limited to, radionuclides such as ^{32}P or ^{35}S, enzymatic labels such as alkaline phosphatase coupled to the probe via avidin/biotin coupling systems, fluorescent labels such as Cy3 and Cy5, and the like. The labeled polynucleotides can be used in Southern or northern analysis, dot blot, or other membrane-based technologies, on chips or other substrates, and in PCR technologies. Hybridization probes are also useful in mapping the naturally occurring genomic sequence. Fluorescent in situ hybridization (FISH) can be correlated with other physical chromosome mapping techniques and genetic map data as described in Heinz-Ulrich et al. (In: Meyers, supra, pp. 965-968). In many cases,

genomic context helps identify genes that encode a particular protein family. (See, e.g., Kirschning et al. (1997) Genomics 46:416-25.)

The polynucleotide can be labeled using standard methods and added to a sample from a subject under conditions for the formation and detection of hybridization complexes. After incubation the sample is washed, and the signal associated with complex formation is quantitated and compared with at least one standard value. Standard values are derived from any control sample, typically one that is free of the suspect disorder and from one that represents a single, specific and preferably, staged disorder. If the amount of signal in the subject sample is distinguishable from the standards, then differential expression in the subject sample indicates the presence of the disorder. Qualitative and quantitative methods for comparing complex formation in subject samples with previously established standards are well known in the art.

Such assays can also be used to evaluate the efficacy of a particular therapeutic treatment regimen in animal studies, in clinical trials, or to monitor the treatment of an individual subject. Once the presence of the disorder has been established and a treatment protocol is initiated, hybridization, amplification, or antibody assays can be repeated on a regular basis to determine when gene or protein expression in the patient begins to approximate that which is observed in a healthy subject. The results obtained from successive assays can be used to show the efficacy of treatment over a period ranging from several hours, e.g. in the case of toxic shock, to many years, e.g. in the case of osteoarthritis.

The polynucleotides can be used on a substrate such as a microarray to monitor gene expression, to identify splice variants, mutations, and polymorphisms. Information derived from analyses of expression patterns can be used to determine gene function, to understand the genetic basis of a disease, to diagnose a disorder, and to develop and monitor the activities of therapeutic agents used to treat a disorder. Microarrays can also be used to detect genetic diversity, single nucleotide polymorphisms, which may characterize a particular population, at the genomic level.

In another embodiment, antibodies or Fabs comprising an antigen binding site that specifically binds the protein can be used for the diagnosis of diseases characterized by the differential expression of the protein. A variety of protocols for measuring protein expression, including ELISAs, RIAs, FACS and antibody arrays, are well known in the art and provide a basis for diagnosing differential or abnormal levels of expression. Standard values for protein expression parallel those reviewed above for nucleotide expression. The amount of complex formation can be quantitated by various methods, preferably by photometric means. Quantities of the protein expressed in subject samples are compared with standard values. Deviation between standard and subject values establishes the parameters for diagnosing or monitoring a particular disorder. Alternatively, one can use

competitive drug screening assays in which neutralizing antibodies capable of binding specifically with the protein compete with a test compound. Antibodies can be used to detect the presence of any peptide which shares one or more epitopes or antigenic determinants with the protein. In one aspect, the antibodies of the present invention can be used for treatment , delivery of therapeutics, or monitoring therapy for pancreatic disorders.

In another aspect, the polynucleotide, or its complement, can be used therapeutically for the purpose of expressing mRNA and protein, or conversely to block transcription or translation of the mRNA. Expression vectors can be constructed using elements from retroviruses, adenoviruses, herpes or vaccinia viruses, or bacterial plasmids, and the like. These vectors can be used for delivery of nucleotide sequences to a particular target cell population, tissue, or organ. Methods well known to those skilled in the art can be used to construct vectors to express the polynucleotides or their complements. (See, e.g., Maulik et al. (1997) Molecular Biotechnology, Therapeutic Applications and Strategies, Wiley-Liss, New York NY.)

Alternatively, the polynucleotide or its complement, can be used for somatic cell or stem cell gene therapy. Vectors can be introduced in vivo, in vitro, and ex vivo. For ex vivo therapy, vectors are introduced into stem cells taken from the subject, and the resulting transgenic cells are clonally propagated for autologous transplant back into that same subject. Delivery of the polynucleotide by transfection, liposome injections, or polycationic amino polymers can be achieved using methods which are well known in the art. (See, e.g., Goldman et al. (1997) Nature Biotechnology 15:462-466.) Additionally, endogenous gene expression can be inactivated using homologous recombination methods which insert an inactive gene sequence into the coding region or other targeted region of the genome. (See, e.g. Thomas et al. (1987) Cell 51: 503-512.)

Vectors containing the polynucleotide can be transformed into a cell or tissue to express a missing protein or to replace a nonfunctional protein. Similarly a vector constructed to express the complement of the polynucleotide can be transformed into a cell to down-regulate protein expression. Complementary or antisense sequences can consist of an oligonucleotide derived from the transcription initiation site; nucleotides between about positions -10 and +10 from the ATG are preferred. Similarly, inhibition can be achieved using triple helix base-pairing methodology. Triple helix pairing is useful because it causes inhibition of the ability of the double helix to open sufficiently for the binding of polymerases, transcription factors, or regulatory molecules. Recent therapeutic advances using triplex DNA have been described in the literature. (See, e.g., Gee et al. In: Huber and Carr (1994) Molecular and Immunologic Approaches, Futura Publishing, Mt. Kisco NY, pp. 163-177.)

Ribozymes, enzymatic RNA molecules, can also be used to catalyze the cleavage of mRNA and decrease the levels of particular mRNAs, such as those comprising the polynucleotides of the invention. (See,

e.g., Rossi (1994) Current Biology 4: 469-471.) Ribozymes can cleave mRNA at specific cleavage sites. Alternatively, ribozymes can cleave mRNAs at locations dictated by flanking regions that form complementary base pairs with the target mRNA. The construction and production of ribozymes is well known in the art and is described in Meyers (supra).

RNA molecules can be modified to increase intracellular stability and half-life. Possible modifications include, but are not limited to, the addition of flanking sequences at the 5' and/or 3' ends of the molecule, or the use of phosphorothioate or 2' O-methyl rather than phosphodiester linkages within the backbone of the molecule. Alternatively, nontraditional bases such as inosine, queosine, and wybutosine, as well as acetyl-, methyl-, thio-, and similarly modified forms of adenine, cytidine, guanine, thymine, and uridine which are not as easily recognized by endogenous endonucleases, can be included.

Further, an antagonist, or an antibody that binds specifically to the protein can be administered to a subject to treat a pancreatic disorder. The antagonist, antibody, or fragment can be used directly to inhibit the activity of the protein or indirectly to deliver a therapeutic agent to cells or tissues which express the protein. The therapeutic agent can be a cytotoxic agent selected from a group including, but not limited to, abrin, ricin, doxorubicin, daunorubicin, taxol, ethidium bromide, mitomycin, etoposide, tenoposide, vincristine, vinblastine, colchicine, dihydroxy anthracin dione, actinomycin D, diphteria toxin, Pseudomonas exotoxin A and 40, radioisotopes, and glucocorticoid.

Antibodies to the protein can be generated using methods that are well known in the art. Such antibodies can include, but are not limited to, polyclonal, monoclonal, chimeric, and single chain antibodies, Fab fragments, and fragments produced by a Fab expression library. Neutralizing antibodies, such as those which inhibit dimer formation, are especially preferred for therapeutic use. Monoclonal antibodies to the protein can be prepared using any technique which provides for the production of antibody molecules by continuous cell lines in culture. These include, but are not limited to, the hybridoma, the human B-cell hybridoma, and the EBV-hybridoma techniques. In addition, techniques developed for the production of chimeric antibodies can be used. (See, e.g., Pound (1998) Immunochemical Protocols, Methods Mol Biol Vol. 80.) Alternatively, techniques described for the production of single chain antibodies can be employed. Fabs which contain specific binding sites for the protein can also be generated. Various immunoassays can be used to identify antibodies having the desired specificity. Numerous protocols for competitive binding or immunoradiometric assays using either polyclonal or monoclonal antibodies with established specificities are well known in the art.

Yet further, an agonist of the protein can be administered to a subject to treat a disorder associated with decreased expression, longevity or activity of the protein.

An additional aspect of the invention relates to the administration of a pharmaceutical or sterile composition, in conjunction with a pharmaceutically acceptable carrier, for any of the therapeutic applications discussed above. Such pharmaceutical compositions can consist of the protein or antibodies, mimetics, agonists, antagonists, or inhibitors of the protein. The compositions can be administered alone or in combination with at least one other agent, such as a stabilizing compound, which can be administered in any sterile, biocompatible pharmaceutical carrier including, but not limited to, saline, buffered saline, dextrose, and water. The compositions can be administered to a subject alone or in combination with other agents, drugs, or hormones.

The pharmaceutical compositions utilized in this invention can be administered by any number of routes including, but not limited to, oral, intravenous, intramuscular, intra-arterial, intramedullary, intrathecal, intraventricular, transdermal, subcutaneous, intraperitoneal, intranasal, enteral, topical, sublingual, or rectal means.

In addition to the active ingredients, these pharmaceutical compositions can contain pharmaceutically-acceptable carriers comprising excipients and auxiliaries which facilitate processing of the active compounds into preparations which can be used pharmaceutically. Further details on techniques for formulation and administration can be found in the latest edition of Remington's Pharmaceutical Sciences (Mack Publishing, Easton PA).

For any compound, the therapeutically effective dose can be estimated initially either in cell culture assays or in animal models such as mice, rats, rabbits, dogs, or pigs. An animal model can also be used to determine the concentration range and route of administration. Such information can then be used to determine useful doses and routes for administration in humans.

A therapeutically effective dose refers to that amount of active ingredient which ameliorates the symptoms or condition. Therapeutic efficacy and toxicity can be determined by standard pharmaceutical procedures in cell cultures or with experimental animals, such as by calculating and contrasting the ED_{50} (the dose therapeutically effective in 50% of the population) and LD_{50} (the dose lethal to 50% of the population) statistics. Any of the therapeutic compositions described above can be applied to any subject in need of such therapy, including, but not limited to, mammals such as dogs, cats, cows, horses, rabbits, monkeys, and most preferably, humans.

Stem Cells and Their Use

SEQ ID NOs:1-13 can be useful in the differentiation of stem cells. Eukaryotic stem cells are able to differentiate into the multiple cell types of various tissues and organs and to play roles in embryogenesis and adult tissue regeneration (Gearhart (1998) Science 282:1061-1062; Watt and Hogan (2000) Science 287:1427-

1430). Depending on their source and developmental stage, stem cells can be totipotent with the potential to create every cell type in an organism and to generate a new organism, pluripotent with the potential to give rise to most cell types and tissues, but not a whole organism; or multipotent cells with the potential to differentiate into a limited number of cell types. Stem cells can be transfected with polynucleotides which can be transiently expressed or can be integrated within the cell as transgenes.

Embryonic stem (ES) cell lines are derived from the inner cell masses of human blastocysts and are pluripotent (Thomson et al. (1998) Science 282:1145-1147). They have normal karyotypes and express high levels of telomerase which prevent senescence and allow the cells to replicate indefinitely. ES cells produce derivatives that give rise to embryonic epidermal, mesodermal and endodermal cells. Embryonic germ (EG) cell lines, which are produced from primordial germ cells isolated from gonadal ridges and mesenteries, also show stem cell behavior (Shamblott et al. (1998) Proc Natl Acad Sci 95:13726-13731). EG cells have normal karyotypes and appear to be pluripotent.

Organ-specific adult stem cells differentiate into the cell types of the tissues from which they were isolated. They maintain their original tissues by replacing cells destroyed from disease or injury. Adult stem cells are multipotent and under proper stimulation can be used to generate cell types of various other tissues (Vogel (2000) Science 287:1418-1419). Hematopoietic stem cells from bone marrow provide not only blood and immune cells, but can also be induced to transdifferentiate to form brain, liver, heart, skeletal muscle and smooth muscle cells. Similarly mesenchymal stem cells can be used to produce bone marrow, cartilage, muscle cells, and some neuron-like cells, and stem cells from muscle have the ability to differentiate into muscle and blood cells (Jackson et al. (1999) Proc Natl Acad Sci 96:14482-14486). Neural stem cells, which produce neurons and glia, can also be induced to differentiate into heart, muscle, liver, intestine, and blood cells (Kuhn and Svendsen (1999) BioEssays 21:625-630); Clarke et al. (2000) Science 288:1660-1663; Gage (2000) Science 287:1433-1438; and Galli et al. (2000) Nature Neurosci 3:986-991).

Neural stem cells can be used to treat neurological disorders such as Alzheimer's disease, Parkinson's disease, and multiple sclerosis and to repair tissue damaged by strokes and spinal cord injuries. Hematopoietic stem cells can be used to restore immune function in immunodeficient patients or to treat autoimmune disorders by replacing autoreactive immune cells with normal cells to treat diseases such as multiple sclerosis, scleroderma, rheumatoid arthritis, and systemic lupus erythematosus. Mesenchymal stem cells can be used to repair tendons or to regenerate cartilage to treat arthritis. Liver stem cells can be used to repair liver damage. Pancreatic stem cells can be used to replace islet cells to treat diabetes. Muscle stem cells can be used to regenerate muscle to treat muscular dystrophies (Fontes and Thomson (1999) BMJ 319:1-3; Weissman (2000)

Science 287:1442-1446 Marshall (2000) Science 287:1419-1421; and Marmont (2000) Ann Rev Med 51:115-134).

EXAMPLES

It is to be understood that this invention is not limited to the particular devices, machines, materials and methods described. Although particular embodiments are described, equivalent embodiments can be used to practice the invention. The described embodiments are provided to illustrate the invention and are not intended to limit the scope of the invention which is limited only by the appended claims.

I cDNA LIBRARY CONSTRUCTION

The cDNA library, PANCNOT05, was selected as an example to demonstrate the construction of the cDNA libraries from which the sequences used to identify genes associated with pancreatic disorders were derived. The PANCNOT05 cDNA library was constructed from cytologically normal pancreas tissue obtained from a 2-year-old Hispanic male who died of cerebral anoxia.

The frozen tissue was homogenized and lysed using a POLYTRON homogenizer (Brinkmann Instruments, Westbury NJ) in guanidinium isothiocyanate solution. The lysate was centrifuged over a 5.7 M CsCl cushion using an SW28 rotor in an L8-70M ultracentrifuge (BeckmanCoulter, Fullerton CA) for 18 hours at 25,000 rpm at ambient temperature. The RNA was extracted with acid phenol, pH 4.0, precipitated using 0.3 M sodium acetate and 2.5 volumes of ethanol, resuspended in RNAse-free water, and DNAse treated at 37C. RNA extraction and precipitation were repeated as before. The mRNA was isolated using the OLIGOTEX kit (Qiagen, Chatsworth CA) and used to construct the cDNA library.

The mRNA was handled according to the recommended protocols in the SUPERSCRIPT plasmid system (Life Technologies). cDNAs were fractionated on a SEPHAROSE CL4B column (Amersham Pharmacia Biotech), and those cDNAs exceeding 400 bp were ligated into pSport I plasmid. The plasmid was subsequently transformed into DH5α competent cells (Life Technologies).

II Isolation and Sequencing of cDNA Clones

Plasmid DNA was released from the bacterial cells and purified using the REAL PREP 96 plasmid kit (Qiagen). This kit enabled the simultaneous purification of 96 samples in a 96-well block using multi-channel reagent dispensers. The recommended protocol was employed except for the following changes: 1) the bacteria were cultured in 1 ml of sterile TERRIFIC BROTH (BD Biosciences, San Jose CA) with carbenicillin at 25 mg/L and glycerol at 0.4%; 2) the cultures were incubated for 19 hours after inoculation and the cells were lysed in 0.3 ml of lysis buffer; and 3) the plasmid DNA pellet was precipitated in isopropanol and then resuspended in 0.1 ml of distilled water. After the last step in the protocol, samples were transferred to a 96-well block for storage at 4C.

The cDNAs were prepared using a MICROLAB 2200 system (Hamilton, Reno NV) in combination with DNA ENGINE thermal cyclers (MJ Research, Watertown MA). The cDNAs were sequenced by the method of Sanger and Coulson (1975; J Mol Biol 94:441-448) using ABI PRISM 377 DNA sequencing systems (Applied Biosystems). Most of the cDNAs were sequenced using standard ABI protocols and kits at solution volumes of 0.25x - 1.0x. In the alternative, some of the cDNAs were sequenced using solutions and dyes from APB.

III SELECTION, ASSEMBLY, AND CHARACTERIZATION OF SEQUENCES

The polynucleotides used for co-expression analysis were assembled from EST sequences, 5' and 3' long read sequences, and full length coding sequences. Of the 41,419 assembled sequences used in the analysis, each was expressed in at least five cDNA libraries.

The assembly process is described as follows. EST sequence chromatograms were processed and verified. Quality scores were obtained using PHRED (Ewing et al. (1998) Genome Res 8:175-185; Ewing and Green (1998) Genome Res 8:186-194), and edited sequences were loaded into a relational database management system (RDBMS). The sequences were clustered using BLAST with a product score of 50. All clusters of two or more sequences created a bin which represents one transcribed gene.

Assembly of the component sequences within each bin was performed using a modification of Phrap, a publicly available program for assembling DNA fragments (Green, P. University of Washington, Seattle WA). Bins that showed 82% identity from a local pair-wise alignment between any of the consensus sequences were merged.

Bins were annotated by screening the consensus sequence in each bin against public databases, such as GBpri and GenPept from NCBI. The annotation process involved a FASTn screen against the GBpri database in GenBank. Those hits with a percent identity of greater than or equal to 75% and an alignment length of greater than or equal to 100 base pairs were recorded as homolog hits. The residual unannotated sequences were screened by FASTx against GenPept. Those hits with an E value of less than or equal to 10^{-8} were recorded as homolog hits.

Sequences were then reclustered using BLASTn and Cross-Match, a program for rapid amino acid and nucleic acid sequence comparison and database search (Green, supra), sequentially. Any BLAST alignment between a sequence and a consensus sequence with a score greater than 150 was realigned using cross-match. The sequence was added to the bin whose consensus sequence gave the highest Smith-Waterman score (Smith et al. (1992) Protein Engineering 5:35-51) amongst local alignments with at least 82% identity. Non-matching sequences were moved into new bins, and assembly processes were repeated.

IV HOMOLOGY SEARCHING OF POLYNUCLEOTIDES AND THEIR ENCODED PROTEINS

The polynucleotides of the Sequence Listing or their encoded proteins were used to query databases such as GenBank, SwissProt, BLOCKS, and the like. These databases that contain previously identified and annotated sequences or domains were searched using BLAST or BLAST 2 (Altschul et al. supra; Altschul, supra) to produce alignments and to determine which sequences were exact matches or homologs. The alignments were to sequences of prokaryotic (bacterial) or eukaryotic (animal, fungal, or plant) origin. Alternatively, algorithms such as the one described in Smith and Smith (1992, Protein Engineering 5:35-51) could have been used to deal with primary sequence patterns and secondary structure gap penalties. All of the sequences disclosed in this application have lengths of at least 49 nucleotides, and no more than 12% uncalled bases (where N is recorded rather than A, C, G, or T).

As detailed in Karlin and Altschul (1993; Proc Natl Acad Sci 90:5873-5877), BLAST matches between a query sequence and a database sequence were evaluated statistically and only reported when they satisfied the threshold of 10^{-25} for nucleotides and 10^{-14} for peptides. Homology was also evaluated by product score calculated as follows: the % nucleotide or amino acid identity [between the query and reference sequences] in BLAST is multiplied by the % maximum possible BLAST score [based on the lengths of query and reference sequences] and then divided by 100. In comparison with hybridization procedures used in the laboratory, the electronic stringency for an exact match was set at 70, and the conservative lower limit for an exact match was set at approximately 40 (with 1-2% error due to uncalled bases).

The BLAST software suite, freely available sequence comparison algorithms (NCBI, Bethesda MD; http://www.ncbi.nlm.nih.gov/gorf/bl2.html), includes various sequence analysis programs including "blastn" that is used to align nucleic acid molecules and BLAST 2 that is used for direct pairwise comparison of either nucleic or amino acid molecules. BLAST programs are commonly used with gap and other parameters set to default settings, e.g.: Matrix: BLOSUM62; Reward for match: 1; Penalty for mismatch: -2; Open Gap: 5 and Extension Gap: 2 penalties; Gap x drop-off: 50; Expect: 10; Word Size: 11; and Filter: on. Identity or similarity is measured over the entire length of a sequence or some smaller portion thereof. Brenner et al. (1998; Proc Natl Acad Sci 95:6073-6078, incorporated herein by reference) analyzed the BLAST for its ability to identify structural homologs by sequence identity and found 30% identity is a reliable threshold for sequence alignments of at least 150 residues and 40%, for alignments of at least 70 residues.

The polynucleotides of this application were compared with assembled consensus sequences or templates found in the LIFESEQ GOLD database. Component sequences from cDNA, extension, full length, and shotgun sequencing projects were subjected to PHRED analysis and assigned a quality score. All sequences with an acceptable quality score were subjected to various pre-processing and editing pathways to

remove low quality 3' ends, vector and linker sequences, polyA tails, Alu repeats, mitochondrial and ribosomal sequences, and bacterial contamination sequences. Edited sequences had to be at least 50 bp in length, and low-information sequences and repetitive elements such as dinucleotide repeats, Alu repeats, and the like, were replaced by "Ns" or masked.

Edited sequences were subjected to assembly procedures in which the sequences were assigned to polynucleotide bins. Each sequence could only belong to one bin, and sequences in each bin were assembled to produce a template. Newly sequenced components were added to existing bins using BLAST and CROSSMATCH. To be added to a bin, the component sequences had to have a BLAST quality score greater than or equal to 150 and an alignment of at least 82% local identity. The sequences in each bin were assembled using PHRAP. Bins with several overlapping component sequences were assembled using DEEP PHRAP. The orientation of each template was determined based on the number and orientation of its component sequences.

Bins were compared to one another and those having local similarity of at least 82% were combined and reassembled. Bins having templates with less than 95% local identity were split. Templates were subjected to analysis by STITCHER/EXON MAPPER algorithms that analyze the probabilities of the presence of splice variants, alternatively spliced exons, splice junctions, differential expression of alternative spliced genes across tissue types or disease states, and the like. Assembly procedures were repeated periodically, and templates were annotated using BLAST against GenBank databases such as GBpri. An exact match was defined as having from 95% local identity over 200 base pairs through 100% local identity over 100 base pairs and a homolog match as having an E-value (or probability score) of $\leq 1 \times 10^{-8}$. The templates were also subjected to frameshift FASTx against GENPEPT, and homolog match was defined as having an E-value of $\leq 1 \times 10^{-8}$. Template analysis and assembly was described in USSN 09/276,534, filed March 25, 1999.

Following assembly, templates were subjected to BLAST, motif, and other functional analyses and categorized in protein hierarchies using methods described in USSN 08/812,290 and USSN 08/811,758, both filed March 6, 1997; in USSN 08/947,845, filed October 9, 1997; and in USSN 09/034,807, filed March 4, 1998. Then templates were analyzed by translating each template in all three forward reading frames and searching each translation against the PFAM database of hidden Markov model-based protein families and domains using the HMMER software package (Washington University School of Medicine, St. Louis MO; http://pfam.wustl.edu/).

The polynucleotide was further analyzed using MACDNASIS PRO software (Hitachi Software Engineering), and LASERGENE software (DNASTAR) and queried against public databases such as the

GenBank rodent, mammalian, vertebrate, prokaryote, and eukaryote databases, SwissProt, BLOCKS, PRINTS, PFAM, and Prosite.

V DESCRIPTION OF GENES KNOWN TO BE ASSOCIATED WITH INSULIN SYNTHESIS

Eight genes known to be associated with insulin synthesis were selected to identify co-expressing novel polynucleotides. They are described below.

Gene	Description & references
Preproinsulin	Precursor for insulin, a peptide hormone synthesized in the beta islet cells of the pancreas. Insulin regulates serum glucose (Darnell et al. (1990) Molecular Cell Biology, WH Freeman, New York NY, p. 743).
Proglucagon	Precursor for glucagon, a peptide hormone synthesized in the pancreas and intestines. Glucagon increases serum glucose levels by inducing the liver to produce and release glucose, thus counter-acting the effects of insulin. (Darnell et al. (supra) p. 743).
Reg	Regenerating (Reg) gene family (Alternate name: lithostathine) whose members include Reg-1 alpha, Reg-1 beta, and Reg-related protein. (Miyashita et al. (1995) FEBS Lett 377:429-33). Reg stimulates growth of the beta islet cells; and its expression is correlated with insulin expression (Baeza et al. (1996) Diabetes Metab 22:229-34). Reg-1 alpha is an effective therapy for diabetes in mice, in combination with the immunoregulator drug linomide. (Gross et al. (1998) Endocrinology 139: 2369-74).
Lipase	Pancreatic lipase expression is elevated in diabetes and restored to normal levels by insulin (Tsai et al. (1994) Am J Physiol 267:G575-83; Sztalryd and Kraemer (1995) Metabolism 44:1391-6).
Colipase	Colipase is a pancreatic exocrine protein whose synthesis increases in diabetic rats; synthesis of colipase is inhibited by insulin (Duan et al. (1991) Pancreas 6:595-602; Duan and Erlanson-Albertsson (1992) Pancreas 7:465-71).
HiAPP	Human islet amyloid polypeptide (HiAPP) is a hormone-like peptide expressed in the insulin-producing beta cells of the endocrine pancreas (Nishi et al. (1989) Mol Endocrinol 3:1775-81).

VI CO-EXPRESSION AMONG KNOWN MARKER GENES AND NOVEL POLYNUCLEOTIDES

The co-expression of the eight known genes, designated 1-8 on both the horizontal and vertical axes, with each other is shown below. The numbers in the table are the negative log of the p-value (- log p) for the co-expression between two genes. For example, reading the values at the intersection of the horizontal and vertical designations for each set, the co-expression between insulin (3) and colipase (2) at a p-value of 17, and between glucagon (7) and colipase (2), at a p-value of 11, are both very highly significant. The fact that co-expression analysis successfully identified the strong associations among the known genes validates the GBA or

co-expression method for identifying polynucleotides that are co-expressed with the known genes. The degree of association was measured by probability values, and the threshold probability used in this analysis was less than 0.0001.

Using the LIFESEQ GOLD database (Incyte Genomics), the method identified novel polynucleotides from among a total of 41,419 assembled polynucleotides that showed highly significant association with the known genes. The process was reiterated until the number of polynucleotides was reduced to the final thirteen polynucleotides shown below. The tabular entries show the p-value (- log p) for the co-expression between each known marker gene and each novel polynucleotide. The novel polynucleotides are identified in the table by their Incyte clone numbers and the known genes their abbreviated names as shown in Example IV above. For each polynucleotide, the p-value is the probability that the observed co-expression is due to chance, using the Fisher Exact Test.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1 Lipase																				
2 Colipase	11																			
3 Insulin	11	17																		
4 Reg-1 beta	5	5	5																	
5 Reg-1 alpha	9	10	12	5																
6 Reg-related	7	6	6	7	6															
7 Glucagon	9	11	16	5	10	6														
8 HiAPP	5	4	4	7	4	6	4													
9 2091133	5	4	4	4	2	4	2													
10 3836037	5	5	5	4	5	4	5	4	4											
11 3833667	5	5	5	4	5	4	5	4	4	7										
12 3664676	3	5	5	0	5	0	5	0	0	2	2									
13 3835361	5	5	5	2	5	2	5	2	2	4	4	4								
14 884692	3	5	5	2	5	2	5	2	2	2	2	4	4							
15 2383628	14	16	16	5	10	7	12	5	5	5	5	3	5	3						
16 888246	7	6	6	4	4	4	4	4	6	7	7	2	4	2	7					
17 2774542	8	7	7	4	7	6	8	4	4	4	4	2	4	2	9	6				
18 888309	5	5	5	4	5	4	5	4	4	7	7	2	4	2	5	7	4			
19 951335	12	11	11	5	10	7	8	4	4	5	5	3	5	3	13	7	8	5		
20 2777115	11	10	10	3	7	3	7	3	5	6	6	3	6	3	12	8	7	6	10	
21 2075919	11	12	12	5	7	7	7	7	7	5	5	3	5	3	12	7	9	5	13	8

The highest co-expression value is obtained when the highest p-value found along the horizontal line following each SEQ ID NO (clone number) is correlated with a known marker gene (numbers 1-8 along the top line of the table). For example, clone number 2383628 (number 15), has a p-value of 14 as it co-expresses with lipase (number 1) and a p-value of 16 as it co-expresses with colipase (number 2); these values greatly exceed the threshold p-value for this experiment and are very highly significant. The data above can be summarized by

reducing it to a single highest co-expression (-log p) value for each intersecting known gene and unknown polynucleotide and naming at least one pancreatic disorder associated with expression of the known gene. The summary table shown below:

Gene	p-value*	SEQ ID	Incyte clone	Pancreatic Disorder	% specificity**
colipase	11	1	223163CT1	type 1 diabetes	77
insulin	5	2	884692CB1	type 1 diabetes	100
lipase	7	3	888246CB1	type 1 diabetes	99
insulin	5	4	888309CB1	type 1 diabetes	100
lipase	12	5	951335CB1	type 1 diabetes	99
HiAPP	6	6	2091133CT1	type 1 diabetes	92
colipase	16	7	2383628CB1	type 1 diabetes	95
glucagon	8	8	2774542CB1	islet cell hyperplasia	47
lipase	11	9	2777115CB1	type 1 diabetes	100
glucagon	5	10	3664676CB1	islet cell hyperplasia	100
insulin	5	11	3833667CB1	type 1 diabetes	96
colipase	5	12	3835361CB1	type 1 diabetes	100
reg-1 alpha	5	13	3836037CB1	cerebral anoxia	97

* p-value (- log p) = 5 is highly significant

** in pancreas as opposed to any other category (Note: all categories are listed in Example IX below).

VII NOVEL POLYNUCLEOTIDES IDENTIFIED USING GBA

Using the method of Walker (supra), thirteen polynucleotides that exhibit strong association, or co-expression, with known genes that regulate, respond to, or participate in insulin synthesis have been identified.

Polynucleotides comprising the nucleic acid sequences of SEQ ID NOs:1-13 of the present invention were first identified as Incyte Clones 223163, 884692, 888246, 888309, 951335, 2091133, 2383628, 2774542, 2777115, 3664676, 3833667, 3835361, and 3836037, respectively; and assembled according to Example III. As described in Example IV, BLAST and other motif searches were performed for each sequence. SEQ ID NOs:1-13 were translated, and sequence identity with known sequences was sought. SEQ ID NOs:14 and 15 of the present invention were encoded by SEQ ID NOs:1 and 8, respectively. SEQ ID NOs:14 and 15 were also analyzed using BLAST and motif search tools, and the results of these analyses are described below.

SEQ ID NO:2 is 924 nucleic acids in length and has about 92% identity from about nucleotide 211 to about nucleotide 923 with a gene that encodes human pancreatic zymogen granule membrane protein, GP-2 (g1244511) and about 96% match from about nucleotide 923 to about nucleotide 594 with a gene that encodes a human zinc finger protein, ZNF133 (g487782). GP-2 is a 75 kDa glycoprotein released from the membrane of mature zymogen granules by an enzymatic mechanism. The C-terminal region of GP-2 exhibit 26 conserved cysteine residues and includes one epidermal growth factor motif. ZNF133 is a protein that belongs to the

human zinc finger Kruppel family and contains a Kruppel-associated box segment. ZNF133 was localized to chromosome 20p11.2 that is close to the deleted region that characterizes Alagille syndrome.

SEQ ID NO:3 is about 845 nucleotides in length; it shows about 80% identity from about nucleotide 560 to about nucleotide 840 with a complete coding sequence for human protamine 1, protamine 2 and transition protein 2 (g642458) and about 86% identity with a gene that encodes TXA2 gene (EP 490410). TXA2 is a unstable arachidonate metabolite that functions as a potent stimulator of platelet aggregation and a constrictor of vascular and respiratory smooth muscle.

SEQ ID NO:7 is 646 nucleotides in length and shows 77% identity from about nucleotide 1 to about nucleotide 402 with a rat mRNA that encodes syncollin, a secretory granule protein that binds to syntaxin in a Ca^{++}-sensitive manner and functions as a regulator of exocytosis in exocrine tissues (g2258437).

SEQ ID NO:12 is 874 nucleotides in length and shows 98% identity from about nucleotide 363 to about nucleotide 873 with a gene that encodes human pancreatic zymogen granule membrane protein, GP-2 mRNA (g1244511). SEQ ID NO:12 also exhibits 99% identity from about nucleotide 432 to about nucleotide 924 with SEQ ID NO:2. Therefore, SEQ ID NO:2 and SEQ ID NO:12 are potential splice variants with related cellular functions.

SEQ ID NO:1 is 1966 nucleotides in length and shows 77% identity from nucleotide 1 to about nucleotide 1930 with an mRNA that encodes a rat uterus-ovary specific trans-membrane protein (g2460315). This uterus-ovary specific rat protein is expressed upon induction by estrogen. SEQ ID NO: 14, an amino acid sequence encoded by SEQ ID NO:1, is 585 amino acid residues in length and shows about 74% identity from about amino acid residue 22 to about amino acid residue 608 with the rat uterus-ovary specific trans-membrane protein (g2460316). SEQ ID NO:14 also exhibits a transmembrane domain encompassing amino acid residues 576 to 593. Motif analysis shows that SEQ ID NO:14 has eight potential N-glycosylation sites at N30, N58, N68, N149, N272, N371, N395, and N420; twelve potential casein kinase II phosphorylation sites at T23, S109, S290, S349, S372, T380, T409, S464, S521, T557, T613, and T632; three N-myristoylation sites at G21, G29, and G39; thirteen potential protein kinase C phosphorylation sites at T45, S70, S132, S255, S280, T308, T328, T442, T468, S521, S527, T589, and T643; and three potential tyrosine kinase phosphorylation sites at Y180, Y415, and Y528.

SEQ ID NO:8 is 1354 nucleotides in length and shows 99% identity with the human mRNA that codes for AQP8 (g2346968), a member of a family of water channel proteins identified from rat testis that contains the conserved transmembrane domains of the major intrinsic protein (MIP) family. SEQ ID NO:15, the amino acid sequence encoded by SEQ ID NO:8, is 255 amino acids in length and shows 100% sequence identity with

AQP8. BLIMPS analysis shows that SEQ ID NO:15 has six conserved amino acid segments that match the conserved transmembrane domains of the MIP family proteins. These segments encompass amino acid residues 30 to 49, 66 to 90, 103 to 122, 154 to 172, 185 to 207, and 222 to 242.

VIII HYBRIDIZATION TECHNOLOGIES AND ANALYSES

Immobilization of Polynucleotides on a Substrate

The polynucleotides are applied to a substrate by one of the following methods. A mixture of polynucleotides is fractionated by gel electrophoresis and transferred to a nylon membrane by capillary transfer. Alternatively, the polynucleotides are individually ligated to a vector and inserted into bacterial host cells to form a library. The polynucleotides are then arranged on a substrate by one of the following methods. In the first method, bacterial cells containing individual clones are robotically picked and arranged on a nylon membrane. The membrane is placed on LB agar containing selective agent (carbenicillin, kanamycin, ampicillin, or chloramphenicol depending on the vector used) and incubated at 37C for 16 hr. The membrane is removed from the agar and consecutively placed colony side up in 10% SDS, denaturing solution (1.5 M NaCl, 0.5 M NaOH), neutralizing solution (1.5 M NaCl, 1 M Tris-HCl, pH 8.0), and twice in 2xSSC for 10 min each. The membrane is then UV irradiated in a STRATALINKER UV-crosslinker (Stratagene).

In the second method, polynucleotides are amplified from bacterial vectors by thirty cycles of PCR using primers complementary to vector sequences flanking the insert. PCR amplification increases a starting concentration of 1-2 ng nucleic acid to a final quantity greater than 5 μg. Amplified nucleic acids from about 400 bp to about 5000 bp in length are purified using SEPHACRYL-400 beads (APB). Purified nucleic acids are arranged on a nylon membrane manually or using a dot/slot blotting manifold and suction device and are immobilized by denaturation, neutralization, and UV irradiation as described above. Purified nucleic acids are robotically arranged and immobilized on polymer-coated glass slides using the procedure described in USPN 5,807,522. Polymer-coated slides are prepared by cleaning glass microscope slides (Corning, Acton MA) by ultrasound in 0.1% SDS and acetone, etching in 4% hydrofluoric acid (VWR Scientific Products, West Chester PA), coating with 0.05% aminopropyl silane (Sigma-Aldrich) in 95% ethanol, and curing in a 110C oven. The slides are washed extensively with distilled water between and after treatments. The nucleic acids are arranged on the slide and then immobilized by exposing the array to UV irradiation using a STRATALINKER UV-crosslinker (Stratagene). Arrays are then washed at room temperature in 0.2% SDS and rinsed three times in distilled water. Non-specific binding sites are blocked by incubation of arrays in 0.2% casein in phosphate buffered saline (PBS; Tropix, Bedford MA) for 30 min at 60C; then the arrays are washed in 0.2% SDS and rinsed in distilled water as before.

Probe Preparation for Membrane Hybridization

Hybridization probes derived from the polynucleotides of the Sequence Listing are employed for screening cDNAs, mRNAs, or genomic DNA in membrane-based hybridizations. Probes are prepared by diluting the polynucleotides to a concentration of 40-50 ng in 45 μl TE buffer, denaturing by heating to 100C for five min, and briefly centrifuging. The denatured polynucleotide is then added to a REDIPRIME tube (APB), gently mixed until blue color is evenly distributed, and briefly centrifuged. Five μl of [^{32}P]dCTP is added to the tube, and the contents are incubated at 37C for 10 min. The labeling reaction is stopped by adding 5 μl of 0.2M EDTA, and probe is purified from unincorporated nucleotides using a PROBEQUANT G-50 microcolumn (APB). The purified probe is heated to 100C for five min, snap cooled for two min on ice, and used in membrane-based hybridizations as described below.

Probe Preparation for Polymer Coated Slide Hybridization

Hybridization probes derived from mRNA isolated from samples are employed for screening polynucleotides of the Sequence Listing in array-based hybridizations. Probe is prepared using the GEMbright kit (Incyte Genomics) by diluting mRNA to a concentration of 200 ng in 9 μl TE buffer and adding 5 μl 5x buffer, 1 μl 0.1 M DTT, 3 μl Cy3 or Cy5 labeling mix, 1 μl RNAse inhibitor, 1 μl reverse transcriptase, and 5 μl 1x yeast control mRNAs. Yeast control mRNAs are synthesized by in vitro transcription from noncoding yeast genomic DNA (W. Lei, unpublished). As quantitative controls, one set of control mRNAs at 0.002 ng, 0.02 ng, 0.2 ng, and 2 ng are diluted into reverse transcription reaction mixture at ratios of 1:100,000, 1:10,000, 1:1000, and 1:100 (w/w) to sample mRNA respectively. To examine mRNA differential expression patterns, a second set of control mRNAs are diluted into reverse transcription reaction mixture at ratios of 1:3, 3:1, 1:10, 10:1, 1:25, and 25:1 (w/w). The reaction mixture is mixed and incubated at 37C for two hr. The reaction mixture is then incubated for 20 min at 85C, and probes are purified using two successive CHROMA SPIN+TE 30 columns (Clontech, Palo Alto CA). Purified probe is ethanol precipitated by diluting probe to 90 μl in DEPC-treated water, adding 2 μl 1mg/ml glycogen, 60 μl 5 M sodium acetate, and 300 μl 100% ethanol. The probe is centrifuged for 20 min at 20,800xg, and the pellet is resuspended in 12 μl resuspension buffer, heated to 65C for five min, and mixed thoroughly. The probe is heated and mixed as before and then stored on ice. Probe is used in high density array-based hybridizations as described below.

Membrane-based Hybridization

Membranes are pre-hybridized in hybridization solution containing 1% Sarkosyl and 1x high phosphate buffer (0.5 M NaCl, 0.1 M Na_2HPO_4, 5 mM EDTA, pH 7) at 55C for two hr. The probe, diluted in 15 ml fresh hybridization solution, is then added to the membrane. The membrane is hybridized with the probe at 55C for 16 hr. Following hybridization, the membrane is washed for 15 min at 25C in 1mM Tris (pH 8.0), 1% Sarkosyl, and

four times for 15 min each at 25C in 1mM Tris (pH 8.0). To detect hybridization complexes, XOMAT-AR film (Eastman Kodak, Rochester NY) is exposed to the membrane overnight at -70C, developed, and examined visually.

Polymer Coated Slide-based Hybridization

Probe is heated to 65C for five min, centrifuged five min at 9400 rpm in a 5415C microcentrifuge (Eppendorf Scientific, Westbury NY), and then 18 μl are aliquoted onto the array surface and covered with a coverslip. The arrays are transferred to a waterproof chamber having a cavity just slightly larger than a microscope slide. The chamber is kept at 100% humidity internally by the addition of 140 μl of 5xSSC in a corner of the chamber. The chamber containing the arrays is incubated for about 6.5 hr at 60C. The arrays are washed for 10 min at 45C in 1xSSC, 0.1% SDS, and three times for 10 min each at 45C in 0.1xSSC, and dried.

Hybridization reactions are performed in absolute or differential hybridization formats. In the absolute hybridization format, probe from one sample is hybridized to array elements, and signals are detected after hybridization complexes form. Signal strength correlates with probe mRNA levels in the sample. In the differential hybridization format, differential expression of a set of genes in two biological samples is analyzed. Probes from the two samples are prepared and labeled with different labeling moieties. A mixture of the two labeled probes is hybridized to the array elements, and signals are examined under conditions in which the emissions from the two different labels are individually detectable. Elements on the array that are hybridized to equal numbers of probes derived from both biological samples give a distinct combined fluorescence (Shalon WO95/35505).

Hybridization complexes are detected with a microscope equipped with an INNOVA 70 mixed gas 10 W laser (Coherent, Santa Clara CA) capable of generating spectral lines at 488 nm for excitation of Cy3 and at 632 nm for excitation of Cy5. The excitation laser light is focused on the array using a 20X microscope objective (Nikon, Melville NY). The slide containing the array is placed on a computer-controlled X-Y stage on the microscope and raster-scanned past the objective with a resolution of 20 micrometers. In the differential hybridization format, the two fluorophores are sequentially excited by the laser. Emitted light is split, based on wavelength, into two photomultiplier tube detectors (PMT R1477, Hamamatsu Photonics Systems, Bridgewater NJ) corresponding to the two fluorophores. Appropriate filters positioned between the array and the photomultiplier tubes are used to filter the signals. The emission maxima of the fluorophores used are 565 nm for Cy3 and 650 nm for Cy5. The sensitivity of the scans is calibrated using the signal intensity generated by the yeast control mRNAs added to the probe mix. A specific location on the array contains a complementary

DNA sequence, allowing the intensity of the signal at that location to be correlated with a weight ratio of hybridizing species of 1:100,000.

The output of the photomultiplier tube is digitized using a 12-bit RTI-835H analog-to-digital (A/D) conversion board (Analog Devices, Norwood MA) installed in an IBM-compatible PC computer. The digitized data are displayed as an image where the signal intensity is mapped using a linear 20-color transformation to a pseudocolor scale ranging from blue (low signal) to red (high signal). The data is also analyzed quantitatively. Where two different fluorophores are excited and measured simultaneously, the data are first corrected for optical crosstalk (due to overlapping emission spectra) between the fluorophores using the emission spectrum for each fluorophore. A grid is superimposed over the fluorescence signal image such that the signal from each spot is centered in each element of the grid. The fluorescence signal within each element is then integrated to obtain a numerical value corresponding to the average intensity of the signal. The software used for signal analysis is the GEMTOOLS program (Incyte Genomics).

IX TRANSCRIPT IMAGING

The transcript image performed using the LIFESEQ GOLD database (Aug00rel, Incyte Genomics) allows assessment of the relative abundance of expressed genes in one or more cDNA libraries. Criteria for transcript imaging include category, number of cDNAs per library, description of the library, and the like

All sequences and cDNA libraries in the LIFESEQ database were categorized by system, organ/tissue and cell type. The categories included cardiovascular system, connective tissue, digestive system, embryonic structures, endocrine system, exocrine glands, female and male reproductive, germ cells, hemic/immune system, liver, musculoskeletal system, nervous system, pancreas, respiratory system, sense organs, skin, stomatognathic system, unclassified/mixed, and the urinary tract. For each category, the number of libraries in which the sequence was expressed were counted and shown over the total number of libraries in that category. In some transcript images, all normalized or pooled libraries, which have high copy number sequences removed prior to processing, and all mixed or pooled tissues, which are considered non-specific in that they contain more than one tissue type or more than one subject's tissue, can be excluded from the analysis. Cell lines and/or fetal tissue data can also be disregarded unless the elucidation of inherited disorders would be furthered by their inclusion in the analysis.

For purposes of example, the transcript image for SEQ ID NO:2 is shown below. No libraries were excluded from the analysis. SEQ ID NO:2 was only expressed in pancreatic tissues, which agrees with the 100% specificity shown in Example VI above, and the transcript image both shows independent confirmation of the results of the co-expression analysis and demonstrates differential expression of SEQ ID NO:2 in type I diabetes. Expression exceeded that of any other diseased pancreas library, including tumor and cytologically

normal tissue, by greater than five-fold.

SEQ ID NO:2 (Category: Pancreas)

Library	cDNAs	Description	Abundance	% Abundance
PANCNOT23	3920	pancreas, type I diabetes, 43F	9	0.2296
PANCNOT17	4037	pancreas, mw/mets neuroendocrine CA of liver, 65F	2	0.0495
PANCNOT16	2812	pancreas, aw/Patau's, fetal, 20wM	1	0.0356
PANCNOT05	6805	pancreas, 2M	2	0.0294
PANCNOT19	3775	pancreas, 8M	1	0.0265
PANCNOT21	3846	pancreas, 8M	1	0.0260

X COMPLEMENTARY MOLECULES

The complement of the novel polynucleotide, from about 5 bp (e.g., a PNA) to about 5000 bp (e.g., the complement of a cDNA insert), are used to detect or inhibit gene expression. These molecules are selected using LASERGENE software (DNASTAR). Detection is described in Example VIII. To inhibit transcription by preventing promoter binding, the complementary molecule is designed to bind to the most unique 5' sequence and includes nucleotides of the 5' UTR upstream of the initiation codon of the open reading frame. Complementary molecules include genomic sequences (such as enhancers or introns) and are used in "triple helix" base pairing to compromise the ability of the double helix to open sufficiently for the binding of polymerases, transcription factors, or regulatory molecules. To inhibit translation, a complementary molecule is designed to prevent ribosomal binding to the mRNA encoding the protein.

Complementary molecules are placed in expression vectors and used to transform a cell line to test efficacy; into an organ, tumor, synovial cavity, or the vascular system for transient or short term therapy; or into a stem cell, zygote, or other reproducing lineage for long term or stable gene therapy. Transient expression lasts for a month or more with a non-replicating vector and for three months or more if appropriate elements for inducing vector replication are used in the transformation/expression system.

Stable transformation of appropriate dividing cells with a vector encoding the complementary molecule produces a transgenic cell line, tissue, or organism (USPN 4,736,866). Those cells that assimilate and replicate sufficient quantities of the vector to allow stable integration also produce enough complementary molecules to compromise or entirely eliminate activity of the polynucleotide encoding the protein.

XI PROTEIN EXPRESSION

Expression and purification of the protein are achieved using either a cell expression system or an insect cell expression system. The pUB6/V5-His vector system (Invitrogen, Carlsbad CA) is used to express protein in CHO cells. The vector contains the selectable bsd gene, multiple cloning sites, the promoter/enhancer sequence from the human ubiquitin C gene, a C-terminal V5 epitope for antibody detection with anti-V5

antibodies, and a C-terminal polyhistidine (6xHis) sequence for rapid purification on PROBOND resin (Invitrogen). Transformed cells are selected on media containing blasticidin.

Spodoptera frugiperda (Sf9) insect cells are infected with recombinant Autographica californica nuclear polyhedrosis virus (baculovirus). The polyhedrin gene is replaced with the polynucleotide by homologous recombination and the polyhedrin promoter drives transcription. The protein is synthesized as a fusion protein with 6xhis which enables purification as described above. Purified protein is used in the following activity and to make antibodies.

XII PRODUCTION OF ANTIBODIES

The protein is purified using polyacrylamide gel electrophoresis and used to immunize mice or rabbits. Antibodies are produced using the protocols below. Alternatively, the amino acid sequence of the expressed protein is analyzed using LASERGENE software (DNASTAR) to determine regions of high antigenicity. An antigenic epitope, usually found near the C-terminus or in a hydrophilic region is selected, synthesized, and used to raise antibodies. Typically, epitopes of about 15 residues in length are produced using an ABI 431A peptide synthesizer (Applied Biosystems) using Fmoc-chemistry and coupled to KLH (Sigma-Aldrich) by reaction with N-maleimidobenzoyl-N-hydroxysuccinimide ester to increase antigenicity.

Rabbits are immunized with the epitope-KLH complex in complete Freund's adjuvant. Immunizations are repeated at intervals thereafter in incomplete Freund's adjuvant. After a minimum of seven weeks for mouse or twelve weeks for rabbit, antisera are drawn and tested for antipeptide activity. Testing involves binding the peptide to plastic, blocking with 1% bovine serum albumin, reacting with rabbit antisera, washing, and reacting with radio-iodinated goat anti-rabbit IgG. Methods well known in the art are used to determine antibody titer and the amount of complex formation.

XIII PURIFICATION OF NATURALLY OCCURRING PROTEIN USING SPECIFIC ANTIBODIES

Naturally occurring or recombinant protein is purified by immunoaffinity chromatography using antibodies which specifically bind the protein. An immunoaffinity column is constructed by covalently coupling the antibody to CNBr-activated SEPHAROSE resin (APB). Media containing the protein is passed over the immunoaffinity column, and the column is washed using high ionic strength buffers in the presence of detergent to allow preferential absorbance of the protein. After coupling, the protein is eluted from the column using a buffer of pH 2-3 or a high concentration of urea or thiocyanate ion to disrupt antibody/protein binding, and the protein is collected.

XIV SCREENING MOLECULES FOR SPECIFIC BINDING USING POLYNUCLEOTIDE OR PROTEIN

The polynucleotide, or fragments thereof, or the protein, or portions thereof, are labeled with ^{32}P-dCTP, Cy3-dCTP, or Cy5-dCTP (APB), or with BIODIPY or FITC (Molecular Probes, Eugene OR), respectively.

Libraries of candidate molecules or compounds previously arranged on a substrate are incubated in the presence of composition, a labeled polynucleotide or protein. After incubation under conditions for either a nucleic acid or amino acid sequence, the substrate is washed, and any position on the substrate retaining label, which indicates specific binding or complex formation, is assayed, and the ligand is identified. Data obtained using different concentrations of the nucleic acid or protein are used to calculate affinity between the labeled nucleic acid or protein and the bound molecule.

XV TWO-HYBRID SCREEN

A yeast two-hybrid system, MATCHMAKER LexA Two-Hybrid system (Clontech Laboratories, Palo Alto CA), is used to screen for peptides that bind the protein of the invention. A polynucleotide encoding the protein is inserted into the multiple cloning site of a pLexA vector, ligated, and transformed into E. coli. cDNA, prepared from mRNA, is inserted into the multiple cloning site of a pB42AD vector, ligated, and transformed into E. coli to construct a cDNA library. The pLexA plasmid and pB42AD-cDNA library constructs are isolated from E. coli and used in a 2:1 ratio to co-transform competent yeast EGY48[p8op-lacZ] cells using a polyethylene glycol/lithium acetate protocol. Transformed yeast cells are plated on synthetic dropout (SD) media lacking histidine (-His), tryptophan (-Trp), and uracil (-Ura), and incubated at 30C until the colonies have grown up and are counted. The colonies are pooled in a minimal volume of 1x TE (pH 7.5), replated on SD/-His/-Leu/-Trp/-Ura media supplemented with 2% galactose (Gal), 1% raffinose (Raf), and 80 mg/ml 5-bromo-4-chloro-3-indolyl β-d-galactopyranoside (X-Gal), and subsequently examined for growth of blue colonies. Interaction between expressed protein and cDNA fusion proteins activates expression of a LEU2 reporter gene in EGY48 and produces colony growth on media lacking leucine (-Leu). Interaction also activates expression of ß-galactosidase from the p8op-lacZ reporter construct that produces blue color in colonies grown on X-Gal.

Positive interactions between expressed protein and cDNA fusion proteins are verified by isolating individual positive colonies and growing them in SD/-Trp/-Ura liquid medium for 1 to 2 days at 30C. A sample of the culture is plated on SD/-Trp/-Ura media and incubated at 30C until colonies appear. The sample is replica-plated on SD/-Trp/-Ura and SD/-His/-Trp/-Ura plates. Colonies that grow on SD containing histidine but not on media lacking histidine have lost the pLexA plasmid. Histidine-requiring colonies are grown on SD/Gal/Raf/X-Gal/-Trp/-Ura, and white colonies are isolated and propagated. The pB42AD-cDNA plasmid, which contains a polynucleotide encoding a protein that physically interacts with the protein, is isolated from the yeast cells and characterized.

All patents and publications mentioned in the specification are incorporated by reference herein. Various modifications and variations of the described method and system of the invention will be apparent to those skilled in the art without departing from the scope and spirit of the invention. Although the invention has

been described in connection with specific preferred embodiments, it should be understood that the invention as claimed should not be unduly limited to such specific embodiments. Indeed, various modifications of the described modes for carrying out the invention that are obvious to those skilled in the field of molecular biology or related fields are intended to be within the scope of the following claims.